

Hierarchical configuration model

Remco van der Hofstad, Johan S.H. van Leeuwaarden, Clara Stegehuis

Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

December 29, 2015

Abstract

We introduce a class of random graphs with a hierarchical community structure, which we call the hierarchical configuration model. On the inter-community level, the graph is a configuration model, and on the intra-community level, every vertex in the configuration model is replaced by a community: a small graph. These communities may have any shape, as long as they are connected. For these hierarchical graphs, we find the size of the largest component, the degree distribution and the clustering coefficient. Furthermore, we determine the conditions under which a giant percolation cluster exists, and find its size.

1 Introduction and model

A characteristic feature of many real-world complex networks is that the degree distribution obeys a power-law. A popular model for such power-law networks is the *configuration model*, a random graph with a prescribed degree distribution [3], but a major shortcoming of this model is that it is locally tree-like – it contains only a few short cycles and the graph next to most vertices is a tree – while a prominent feature of complex networks is that they often have a community structure [6]. The communities are highly connected and contain many short cycles, while edges between different communities are more scarce. Therefore, the configuration model is not a realistic model to study networks with a community structure.

Several other random graph models have been proposed to include community structures or short cycles. For example, in [1, 4, 18] communities are introduced in the form of households, i.e., complete graphs of a certain size. The random graph then remains a configuration model on the macroscopic level, while on the microscopic level each vertex of the graph can be replaced by a household. The households introduce a community structure in the graph, which creates random graphs with not only a prescribed degree distribution but also a tunable clustering coefficient. Another way to incorporate short cycles in the configuration model is to introduce triangles [14]. In this case, besides the vertex degrees, also the number of triangles each vertex belongs to is prescribed. Then triangles are formed by joining three nodes at random, and regular edges are formed as in the configuration model. The model was extended in [13] to include arbitrary subgraphs. Yet another method to include clustering in random graph models is to use *random intersection graphs* [5], which allows one to prescribe the exponent of the power-law degree sequence of the graph and the amount of clustering. In [16], an algorithm to match the community structure in real-world networks is proposed. This algorithm first randomizes the edges between different communities according to a configuration model, and then it randomizes the edges inside the communities. This creates a graph with a similar community structure and degree distribution as real-world networks. However, this model was only studied through extensive simulations; the analytical properties of this model were not studied in [16].

In this paper, we introduce the hierarchical configuration model, a random graph model that can describe any kind of community structure. Like in [1] and [4], we study a random graph which

is a configuration model on the macroscopic level, and then add communities on the microscopic level. In real-world networks, however, communities do not have to be complete graphs, especially when the communities are large, as is typically observed when using community detection algorithms [7]. We generalize the setting of [1] and [4] to a configuration model in which a vertex can be replaced by any *small* graph. This generalization makes it possible to apply the hierarchical configuration model to real-world data sets. When the community structure of a real-world network is extracted by a community detection algorithm, the hierarchical configuration model is able to produce random graphs that have a similar community structure. Furthermore, due to the general community structure of the hierarchical configuration model, several existing random graph models turn out to be special cases.

This paper is organized as follows. In Section 1.1 we define the hierarchical configuration model. Section 2 presents several analytical results for the hierarchical configuration model, including the condition for a giant component to emerge, the degree distribution and the clustering coefficient. In Section 3 we study bond percolation on the hierarchical configuration model. Section 4 describes examples of graph models in the literature that fit into our framework. Then we show in Section 5 how some stylized community structures affect percolation. Finally, we present some conclusions in Section 6.

Notation We use \xrightarrow{d} for convergence in distribution, and $\xrightarrow{\mathbb{P}}$ for convergence in probability. We say that a sequence of events \mathcal{E}_n happens with high probability (w.h.p.) if $\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{E}_n) = 1$. Furthermore, we write $f(n) = o(g(n))$ if $\lim_{n \rightarrow \infty} f(n)/g(n) = 0$, and $f(n) = O(g(n))$ if $|f(n)|/g(n)$ is uniformly bounded, where $(g(n))_{n \geq 1}$ is nonnegative. We say that $X_n = O_{\mathbb{P}}(b_n)$ for a sequence of random variables $(X_n)_{n \geq 1}$ if $|X_n|/b_n$ is a tight sequence of random variables, and $X_n = o_{\mathbb{P}}(b_n)$ if $X_n/b_n \xrightarrow{\mathbb{P}} 0$.

1.1 Hierarchical configuration model

We now describe the random graph model that we introduce and study in this paper. Consider a random graph G with n communities. A community H is represented by $H = (F, (d_v^{(b)})_{v \in V})$, where $F = (V_F, E_F)$ is a simple, connected graph, and $d_v^{(b)}$ is the number of edges from $v \in V_F$ to other communities. Thus $(d_v^{(b)})_{v \in V_F}$ describes the degrees between the communities. We call $d_v^{(b)}$ the *outside degree* of a vertex. A vertex inside a community also has an *inside degree* $d_v^{(c)}$: the number of edges from that vertex to other vertices in the same community. The sum of the inside- and the outside degree of the vertex is the degree of the vertex, i.e., $d_v = d_v^{(b)} + d_v^{(c)}$. Let $d_H = \sum_{v \in V_F} d_v^{(b)}$ be the total number of edges out of community H . On the macroscopic level, G is a configuration model with degrees d_H . Let this macroscopic configuration model be denoted by $\phi(G)$.

Let $H_n = (F_n, \mathbf{d}_n)$ denote a uniformly chosen community in $[n] = \{1, 2, \dots, n\}$. Furthermore, denote the number of communities of type H in a graph with n communities by $n_H^{(n)}$. Then $n_H^{(n)}/n$ is the fraction of communities that are of type H . Let D_n be the number of outgoing edges from a uniformly chosen community, i.e., $D_n = d_{H_n}$. Let the size of community i be denoted by s_i , and the size of a uniformly chosen community in $[n]$ by S_n . Then the total number of vertices in the graph is $N = \sum_{i=1}^n s_i = n\mathbb{E}[S_n]$. We assume that the following regularity conditions hold:

Condition 1 (Community regularity).

- (i) $P_n(H) = n_H^{(n)}/n \xrightarrow{\mathbb{P}} P(H)$, where $P(H)$ is a probability distribution.
- (ii) $\lim_{n \rightarrow \infty} \mathbb{E}[S_n] = \mathbb{E}[S] < \infty$.

Condition 2 (Intercommunity connectivity).

- (i) $\lim_{n \rightarrow \infty} \mathbb{E}[D_n] = \mathbb{E}[D] < \infty$,
- (ii) $\mathbb{P}(D = 2) < 1$.

Condition 1(i) implies $(F_n, \mathbf{d}_n) \xrightarrow{d} (F, \mathbf{d})$, $D_n \xrightarrow{d} D$ and $S_n \xrightarrow{d} S$. Define

$$p_{k,s}^{(n)} = \sum_{H=(F,\mathbf{d}):|F|=s,d_H=k} P_n(H) \quad (1)$$

$$p_{k,s} = \sum_{H=(F,\mathbf{d}):|F|=s,d_H=k} P(H) \quad (2)$$

as the probability that a uniformly chosen community has size s and outside degree k . Then Condition 1 implies that $p_{k,s}^{(n)} \rightarrow p_{k,s}$ for every (k, s) .

We can think of $P_n(H)$ as the probability that a uniformly chosen community has a certain shape. In a data set we can approximate $P_n(H)$ and use the hierarchical configuration model in the following way. Suppose a community detection algorithm gives the empirical distribution of the community shapes $P_n(H)$. Now we construct a random graph in the way that was described above. The probability that a certain community is of shape H is $P_n(H)$. We condition on the total outside degree to be even so that edges between communities can be formed as in a configuration model. This results in a graph with roughly the same degree sequence as the original graph. Additionally, the community structure in the random graph is the same as in the original graph. This construction preserves more of the microscopic features of the original graph than a standard configuration model with the same degree sequence as the original graph. It also shows the necessity of extending the work of [4, 18] to go beyond the assumption that communities are complete graphs, because communities in real-world networks can be non-complete. Using this construction, the hierarchical configuration model can match the community structure in virtually all complex networks.

2 Model properties

For a connected component of G , we can either count the number of communities in the component, or the number of vertices in it. We denote the number of communities in a component \mathcal{C} by $v(\mathcal{C}^H)$, and the number of communities with outside degree k by $v_k(\mathcal{C}^H)$. The number of vertices in component \mathcal{C} is denoted by $v(\mathcal{C})$. Furthermore, define ν_D as

$$\nu_D = \frac{\mathbb{E}[D(D-1)]}{\mathbb{E}[D]}, \quad (3)$$

where D is the asymptotic community degree of Condition 2. Let $p_k = \mathbb{P}(D = k)$ and let $g(x) = \sum_k p_k x^k$ be the probability generating function of D .

2.1 Giant component

In the standard configuration model, a giant component exists w.h.p. if $\nu_D > 1$ [12]. In the hierarchical configuration model a similar statement holds:

Theorem 3. *Let G be a hierarchical configuration model satisfying Conditions 1 and 2. Let \mathcal{C}_{\max} and \mathcal{C}_2 be its largest and second largest components, respectively. Then,*

(i) *If $\nu_D > 1$,*

$$\frac{v(\mathcal{C}_{\max})}{N} \xrightarrow{\mathbb{P}} \frac{\sum_{k,s} s p_{k,s} (1 - \xi^k)}{\mathbb{E}[S]} > 0, \quad (4)$$

where ξ is the unique solution in $[0, 1)$ of $g'(\xi) = \xi \mathbb{E}[D]$. Furthermore, $v(\mathcal{C}_2)/N \xrightarrow{\mathbb{P}} 0$.

(ii) *If $\nu_D \leq 1$, then $v(\mathcal{C}_{\max})/N \xrightarrow{\mathbb{P}} 0$.*

Proof. Suppose $\nu_D > 1$. By [9, Theorem 4.1], if Condition 2 holds, $D_n \xrightarrow{\mathbb{P}} D$ and $\nu_D > 1$ in a standard configuration model, then w.h.p. there will be one component with a positive fraction of vertices as $n \rightarrow \infty$. Furthermore, the number of vertices in the largest component in a standard configuration model $v(\mathcal{C}_{\max}^{\text{CM}})$ and the number of vertices of degree k in the largest component, $v_k(\mathcal{C}_{\max}^{\text{CM}})$ satisfy

$$v(\mathcal{C}_{\max}^{\text{CM}})/n \xrightarrow{\mathbb{P}} 1 - g(\xi) > 0, \quad (5)$$

$$v_k(\mathcal{C}_{\max}^{\text{CM}})/n \xrightarrow{\mathbb{P}} p_k(1 - \xi^k). \quad (6)$$

If $\nu_D < 1$, then $v(\mathcal{C}_{\max}^{\text{CM}})/n \xrightarrow{\mathbb{P}} 0$. Therefore, if Conditions 1 and 2 hold and $\nu_D > 1$ in the hierarchical configuration model, then there is a component with a positive fraction of the communities as $n \rightarrow \infty$. Hence, we need to prove that the largest hierarchical component is indeed a large component with size given by (4) if $\nu_D > 1$, and that a small hierarchical component is also a small component of G .

We denote the number of communities in the largest hierarchical component with outside degree k and size s by $v_{k,s}(\mathcal{C}_{\max}^{\text{H}})$. Since G is a configuration model on the community level, (5) and (6) apply on the community level. Furthermore, given a community in the largest hierarchical component of outside degree k , its size is *independent* of being in the largest hierarchical component. Moreover, $\sum_k s v_{k,s}(\mathcal{C}_{\max}^{\text{H}})/n \leq \sum_k s p_{k,s}^{(n)}$. Therefore, by Condition 1 the fraction of vertices in the largest hierarchical component satisfies

$$\frac{v(\mathcal{C}_{\max})}{N} = \frac{\sum_{i \in \mathcal{C}_{\max}^{\text{H}}} s_i}{\sum_i s_i} = \frac{\sum_{k,s} n^{-1} s v_{k,s}(\mathcal{C}_{\max}^{\text{H}})}{n^{-1} \sum_i s_i} \xrightarrow{\mathbb{P}} \frac{\sum_{k,s} s p_{k,s}(1 - \xi^k)}{\mathbb{E}[S]} > 0. \quad (7)$$

The last inequality follows from Condition 1(ii) and the fact that $\xi \in [0, 1)$ and $s \geq 1$. Now we need to prove that the largest hierarchical component indeed is the largest component of G . We show that a hierarchical component of size $o_{\mathbb{P}}(n)$ is w.h.p. a component of size $o_{\mathbb{P}}(N)$. Take a hierarchical component \mathcal{C} which is not the largest hierarchical component, so that it is of size $o_{\mathbb{P}}(n)$. Then,

$$\begin{aligned} \frac{v(\mathcal{C})}{N} &= \frac{n^{-1} \sum_{k,s} s v_{k,s}(\mathcal{C}^{\text{H}})}{\mathbb{E}[S_n]} = \frac{n^{-1} \sum_{s=1}^K \sum_k s v_{k,s}(\mathcal{C}^{\text{H}})}{\mathbb{E}[S_n]} + \frac{n^{-1} \sum_{s>K} \sum_k s v_{k,s}(\mathcal{C}^{\text{H}})}{\mathbb{E}[S_n]} \\ &\leq K \frac{n^{-1} \sum_{s=1}^K \sum_k s v_{k,s}(\mathcal{C}^{\text{H}})}{\mathbb{E}[S_n]} + \frac{\mathbb{E}[S_n \mathbb{1}_{\{S_n > K\}}]}{\mathbb{E}[S_n]} \leq K \frac{n^{-1} v(\mathcal{C}^{\text{H}})}{\mathbb{E}[S_n]} + \frac{\mathbb{E}[S_n \mathbb{1}_{\{S_n > K\}}]}{\mathbb{E}[S_n]}. \end{aligned} \quad (8)$$

First we take the limit for $n \rightarrow \infty$, and then we let $K \rightarrow \infty$. By [12], $v(\mathcal{C}^{\text{H}})/n \xrightarrow{\mathbb{P}} 0$, hence the first term tends to zero as $n \rightarrow \infty$. Furthermore, $\mathbb{E}[S_n \mathbb{1}_{\{S_n > K\}}] \rightarrow \mathbb{E}[S \mathbb{1}_{\{S > K\}}]$ as $n \rightarrow \infty$ by Condition 1. By Condition 1(ii), this goes to zero as $K \rightarrow \infty$. Thus, $v(\mathcal{C})/N \xrightarrow{\mathbb{P}} 0$. This proves that the largest hierarchical component is indeed the largest component of G . This also proves (ii), since by [9, Theorem 4.1], if $\nu_D \leq 1$, $v(\mathcal{C}_{\max}^{\text{H}}) = o_{\mathbb{P}}(n)$, so that $v(\mathcal{C}_{\max}) = o_{\mathbb{P}}(N)$. \square

We conclude that if Conditions 1 and 2 hold and $\nu_D > 1$, then a giant component exists in the hierarchical configuration model. Equation (4) gives the fraction of vertices in the largest component. The fraction of vertices in the giant component may be different from the fraction of communities in the giant hierarchical component. If the sizes and the outside degrees of the communities are independent, then the fraction of vertices in the largest component is equal to the fraction of communities in the largest hierarchical component.

Corollary 4. *Suppose that in the hierarchical configuration model G satisfying Conditions 1 and 2, the size of the communities and the outside degrees of the communities are independent. Then, if $\nu_D > 1$,*

$$\frac{v(\mathcal{C}_{\max})}{N} \xrightarrow{\mathbb{P}} 1 - g(\xi), \quad \frac{v(\mathcal{C}_{\max}^{\text{H}})}{n} \xrightarrow{\mathbb{P}} 1 - g(\xi), \quad (9)$$

where ξ is the unique solution in $[0, 1)$ of $g'(\xi) = \xi \mathbb{E}[D]$. Hence the fraction of vertices in the largest component is equal to the fraction of communities in the largest hierarchical component. If the size and the outside degrees are dependent, then this does not have to be true.

Proof. The second equality in (9) is given by [9, Theorem 4.1]. The first equality in (9) follows by substituting $p_{k,s} = p_k p_s$ in (4), so that

$$\frac{v(\mathcal{C}_{\max})}{N} \xrightarrow{\mathbb{P}} \frac{\sum_s s p_s \sum_k p_k (1 - \xi^k)}{\mathbb{E}[S]} = \frac{\mathbb{E}[S](1 - \sum_k p_k \xi^k)}{\mathbb{E}[S]} = 1 - g(\xi). \quad (10)$$

To show that (9) may not hold when the outside degrees and the sizes are dependent, consider the hierarchical configuration model with

$$p_{k,s} = \begin{cases} \frac{1}{3} & \text{if } (k, s) = (3, 10), \\ \frac{2}{3} & \text{if } (k, s) = (1, 1). \end{cases} \quad (11)$$

Since $\nu_D = \frac{6}{5} > 1$, a giant component exists w.h.p. Furthermore, ξ solves

$$\frac{2}{3} + \frac{2}{3}\xi^2 = \frac{5}{3}\xi, \quad (12)$$

which has $\frac{1}{2}$ as its only solution in $[0, 1)$. Therefore, the fraction of communities in the largest component is given by $1 - g(\frac{1}{2}) = \frac{5}{8}$. To find the fraction of vertices in the largest component, we use (4), which gives

$$\frac{1}{4}(\frac{2}{3}(1 - \frac{1}{2}) + 10\frac{1}{3}(1 - (\frac{1}{2})^3)) = \frac{13}{16} > \frac{5}{8}. \quad (13)$$

Thus, the fraction of vertices in the largest component is larger than the fraction of communities in the largest component. \square

If there is a difference between the fraction of communities and the fraction of vertices in the largest component, then this difference is caused by the dependence of the sizes and the outside degrees of the communities. A community with a large outside degree has a higher probability of being in the largest hierarchical component than a community with a small outside degree. In the example in the proof of Corollary 4, the communities with large outside degrees are large communities. This causes the fraction of vertices in the largest component to be larger than the fraction of communities in the largest hierarchical component.

2.2 Degree distribution

In the hierarchical configuration model, the macroscopic configuration model has a fixed degree sequence. The degree distribution of G depends on the sizes and shapes of the communities. Let $n_k^{(H)}$ denote the number of vertices in community H with the sum of their inside degree and outside degree equal to k . Then the degree distribution of the total graph G is described in Proposition 5:

Proposition 5. *Let G be a hierarchical configuration model such that Conditions 1 and 2 hold. The asymptotic probability \hat{p}_k that a randomly chosen vertex inside G has degree k , as $n \rightarrow \infty$ satisfies*

$$\hat{p}_k = \frac{\sum_H P(H) n_k^{(H)}}{\mathbb{E}[S]}. \quad (14)$$

Proof. Consider a hierarchical configuration model G on n communities. Let $n_H^{(n)}$ be the number of communities in G of type H . The total number of vertices of degree k is the sum of the number of degree k vertices inside all communities, hence it equals $\sum_H n_H^{(n)} n_k^{(H)}$. Furthermore, $P_n(H) n_k^{(H)} \leq P_n(H) s_H$, so that $\lim_{n \rightarrow \infty} \sum_H P_n(H) n_k^{(H)} = \sum_H P(H) n_k^{(H)}$ by Condition 1. This gives

$$\hat{p}_k^{(n)} = \frac{n^{-1} \sum_H n_H n_k^{(H)}}{n^{-1} N} = \frac{\sum_H P_n(H) n_k^{(H)}}{\mathbb{E}[S_n]} \xrightarrow{\mathbb{P}} \hat{p}_k, \quad (15)$$

as $n \rightarrow \infty$. \square

2.2.1 Power-law shift in dense communities

Proposition 5 shows that the outside degree distribution and the degree distribution of the graph may be different. A case of special interest is when the degree distribution follows a power-law. Let $F_{X_n}(k)$ be the empirical distribution function of n observations. Then we say that X_n follows a power law with exponent τ if $0 < c_1 < c_2$ and k_n exist such that

$$1 - F_{X_n}(k) \geq c_1 k^{-\tau+1} \quad \forall k \leq k_n, \quad (16)$$

$$1 - F_{X_n}(k) \leq c_2 k^{-\tau+1} \quad \forall k, \quad (17)$$

where $k_n \rightarrow \infty$ as $n \rightarrow \infty$.

In the case of a power-law degree distribution, the degree distribution and the community size distribution can be related when the communities are dense enough. We call a community H (η, ε) -dense if

$$\#\{v \in H : d_v^{(c)} \geq \eta s\} \geq \varepsilon s. \quad (18)$$

This condition states that in every community, at least a fraction of ε vertices have edges to at least a fraction of η other vertices of the same community. The special case in which $\varepsilon = \eta = 1$ corresponds to complete graph communities. We now show that for dense communities, the power-law exponents of the community sizes and the degree distribution are related. For the special case of household communities, this relation was already observed in [18].

Proposition 6. *Let G be a hierarchical configuration model such that Conditions 1 and 2 hold. Suppose that there exists a $K \geq 0$ such that $d^{(b)} \leq K s_i$ for all vertices, where s_i is the community size. Furthermore, assume that there exist $\varepsilon, \eta > 0$ such that every community of G is (η, ε) -dense. Then, the community size distribution S follows a power-law distribution with exponent τ' with $\tau' > 2$ if and only if the degree distribution follows a power-law with exponent $\tau = \tau' - 1$ where $\tau > 1$.*

Proof. First, assume that S_n obeys a power-law with exponent τ' , so that for some $0 < b_1 < b_2$ and k_n , $b_1 k^{-\tau'+1} \leq 1 - F_{S_n}(k)$ for all $k \leq k_n$ and $1 - F_{S_n}(k) \leq b_2 k^{-\tau'+1}$ for all k . Then the cumulative distribution function of the degrees \hat{D}_N in a hierarchical configuration model G on N vertices, $F_{\hat{D}_N}(k)$, satisfies

$$\begin{aligned} 1 - F_{\hat{D}_N}(k) &= \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{d_i \geq k\} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{d_i^{(c)} + d_i^{(b)} \geq k\} \geq \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{d_i^{(c)} \geq k\} \\ &\geq \frac{\varepsilon}{N} \sum_{i=1}^N \mathbb{1}\{\eta s_i \geq k\} = \frac{\varepsilon}{n \mathbb{E}[S_n]} \sum_{i=1}^n s_i \mathbb{1}\{s_i \geq \frac{k}{\eta}\} \geq \frac{\varepsilon}{n \mathbb{E}[S_n]} \sum_{i=1}^n \frac{k}{\eta} \mathbb{1}\{s_i \geq \frac{k}{\eta}\} \\ &= \frac{k\varepsilon}{\eta} (1 - F_{S_n}(k/\eta)) \geq \varepsilon c_1(\eta) k^{-\tau'+2} = \varepsilon c_1(\eta) k^{-\tau+1} \quad \forall k \leq k_n \eta, \end{aligned} \quad (19)$$

where $c_1(\eta)$ is a constant depending on η . Because the communities are simple, $d_i^{(c)} \leq s_i - 1$ for all vertices. Hence,

$$\begin{aligned} 1 - F_{\hat{D}_N}(k) &= \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{d_i^{(c)} + d_i^{(b)} \geq k\} \leq \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{(s_i - 1) + s_i K \geq k\} \\ &= \frac{1}{n \mathbb{E}[S_n]} \sum_{i=1}^n s_i \mathbb{1}\{s_i \geq \frac{k+1}{K+1}\} = \frac{1}{\mathbb{E}[S_n]} \mathbb{E}[S_n \mathbb{1}\{S_n \geq \frac{k+1}{K+1}\}] \\ &= \frac{1}{\mathbb{E}[S_n]} \left(\frac{k+1}{K+1} \left(1 - F_{S_n} \left(\frac{k+1}{K+1} \right) \right) + \sum_{j \geq (k+1)/(K+1)} (1 - F_{S_n}(j)) \right) \\ &\leq c_2(K) k^{-\tau'+2} = c_2(K) k^{-\tau+1}, \end{aligned} \quad (20)$$

for all k , where $c_2(K)$ is a constant depending on K . Taking the limit of $N \rightarrow \infty$, equation (19) and (20) imply that the degree distribution of G follows a power-law with exponent τ , which proves the first part.

Now assume that the degree distribution \hat{D}_N of G obeys a power-law with exponent τ , so that $0 < b_3 < b_4$ and k_N exist such that (16) and (17) are satisfied. Then the community sizes are minimized if each community is a complete graph, and each vertex has exactly Ks_i half-edges. Then, $d_v = (s_i - 1) + Ks_i$ or $s_i = (d_v + 1)/(K + 1)$. Hence, the cumulative distribution of S in a graph with n communities, F_{S_n} satisfies

$$\begin{aligned}
1 - F_{S_n}(k) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{s_i \geq k\} \geq \frac{\mathbb{E}[S_n]}{N} \sum_{i=1}^N \frac{K+1}{d_i+1} \mathbb{1}\left\{\frac{d_i+1}{K+1} \geq k\right\} \\
&= \frac{\mathbb{E}[S_n]}{N} \sum_{i=1}^N \frac{K+1}{d_i+1} \mathbb{1}\{d_i \geq k(K+1) - 1\} = \mathbb{E}[S_n](K+1) \mathbb{E}\left[\frac{\mathbb{1}\{\hat{D}_N \geq k(K+1) - 1\}}{\hat{D}_N + 1}\right] \\
&= \mathbb{E}[S_n](K+1) \mathbb{P}(\hat{D}_N > k(K+1) - 1) \mathbb{E}[1/(\hat{D}_N + 1) \mid \hat{D}_N \geq k(K+1) - 1] \\
&\geq \mathbb{E}[S_n](K+1) \frac{\mathbb{P}(\hat{D}_N > k(K+1) - 1)}{\mathbb{E}[\hat{D}_N + 1 \mid \hat{D}_N \geq k(K+1) - 1]} \\
&= \mathbb{E}[S_n](K+1) \frac{\mathbb{P}(\hat{D}_N > k(K+1) - 1)^2}{\mathbb{E}[(\hat{D}_N + 1) \mathbb{1}\{\hat{D}_N + 1 \geq k(K+1)\}]}.
\end{aligned} \tag{21}$$

The term in the denominator satisfies

$$\begin{aligned}
\mathbb{E}[(\hat{D}_N + 1) \mathbb{1}\{\hat{D}_N + 1 \geq k(K+1)\}] &\leq k(K+1) \mathbb{P}(\hat{D}_N + 1 > k(K+1)) + \sum_{j \geq k(K+1)} \mathbb{P}(\hat{D}_N + 1 \geq j) \\
&\leq C(K) b_4 (k(K+1))^{-\tau+2},
\end{aligned} \tag{22}$$

for all k , where $C(K)$ is a constant depending on K . Combining (21) and (22) yields

$$1 - F_{S_n}(k) \geq \mathbb{E}[S_n](K+1) C(K) \frac{b_3 (k(K+1))^{-2(\tau+1)}}{b_4 (k(K+1))^{-\tau+2}} \geq c_3(K) k^{-\tau} = c_3(K) k^{-\tau'+1}, \tag{23}$$

for all $k \leq k_N/(K+1)$, where $c_3(K)$ is a constant depending on K . We also have

$$\begin{aligned}
1 - F_{S_n}(k) &= \frac{\mathbb{E}[S_n]}{N} \sum_{i=1}^N \frac{1}{s_i} \mathbb{1}\{s_i \geq k\} \leq \frac{\mathbb{E}[S_n]}{N} \sum_{i=1}^N \frac{1}{k} \mathbb{1}\{s_i \geq k\} \leq \frac{\mathbb{E}[S_n]}{\varepsilon N} \sum_{i=1}^N \frac{1}{k} \mathbb{1}\{d_i^{(c)} \geq k\eta\} \\
&\leq \frac{\mathbb{E}[S_n]}{\varepsilon N} \sum_{i=1}^N \frac{1}{k} \mathbb{1}\{d_i^{(c)} + d_i^{(b)} \geq k\eta\} \leq \frac{1}{\varepsilon} c_4(\eta) k^{-\tau} = \frac{1}{\varepsilon} c_4(\eta) k^{-\tau'+1},
\end{aligned} \tag{24}$$

for all k . Taking the limit of $n \rightarrow \infty$ proves that S has a power-law distribution with exponent $\tau' = \tau + 1$. \square

Proposition 6 relates the degree distribution to the community size distribution of the hierarchical configuration model. Under a more strict assumption on the outside degrees of individual vertices, this also establishes a similar relation between the degree distribution of G and the outside degree distribution of the communities in case of a power-law degree distribution:

Corollary 7. *Let G be a hierarchical configuration model satisfying Conditions 1 and 2. Suppose that there exists a $K \geq 0$ such that $d^{(c)} \leq K$ for all vertices. Furthermore, assume that there exist $\varepsilon, \eta > 0$ such that every community of G is (η, ε) -dense. Then, the outside degree distribution of G cannot have a power-law distribution with exponent smaller than $\tau + 1$ if the degree distribution of G follows a power-law distribution with exponent τ , where $\tau > 1$.*

Proof. By Proposition 6, S follows a power-law with exponent $\tau+1$. Since $d_v^{(c)} \leq K$, also $d_{H_i} \leq Ks_i$ for all communities H_i . Therefore, $D \preceq KS$, and hence D cannot have a power-law distribution with exponent smaller than $\tau + 1$. \square

Corollary 7 shows that the degrees between communities are smaller than the degrees of the graph. This is consistent with our view of communities being highly connected, while edges between communities are more scarce. For example, if the degree distribution follows a power-law with $\tau \in (2, 3)$, then the outside degree distribution has an exponent of at least 3. Therefore, the inter-community degree distribution has finite variance, whereas the degree distribution has infinite variance. A property of configuration models with power-law exponents in $(2, 3)$ is that the probability of obtaining a simple graph vanishes. However, the inter-community connections in the hierarchical configuration model have exponent larger than 3, so that the probability of obtaining a simple graph remains uniformly positive. This suggests that the hierarchical configuration model is able to produce a random graph which has a positive probability of being simple, while the degree distribution has an exponent in $(2, 3)$.

In a companion paper [17], we study these power-law relations in more detail, and we show that in the case of communities that are not dense, different relations between τ and τ' may hold.

2.2.2 The probability of obtaining a simple graph

In the standard configuration model, the probability of obtaining a simple graph converges to $e^{-\nu/2 - \nu^2/4}$ under the condition that $\mathbb{E}[D^2] < \infty$ [8]. In the hierarchical configuration model, the probability of obtaining a simple graph is largely dependent on the shapes of the communities. Since we assumed that the communities are simple, only the inter-community edges can create self-loops and multiple edges.

Suppose that each vertex in a community has at most one half-edge to other communities, $d_v^{(b)} \in \{0, 1\}$. A double edge in the macroscopic configuration model corresponds to a community where two vertices have an edge to the same other community. Since $d_v^{(b)} \in \{0, 1\}$, a double edge in the macroscopic configuration model cannot correspond to a double edge in the hierarchical configuration model. A self-loop in the macroscopic configuration model corresponds to an edge from one vertex v inside a community to another vertex w inside the same community. This self-loop in the macroscopic configuration model corresponds to a double edge in the hierarchical configuration model if an edge from v to w was already present in the community. Thus, when $d_v^{(b)} \in \{0, 1\}$ the probability that the macroscopic configuration model is simple is lower bounded by the probability that no self-loops exist in the macroscopic configuration model,

$$\liminf_{n \rightarrow \infty} \mathbb{P}(G_n \text{ simple}) \geq e^{-\nu_D/2}. \quad (25)$$

In the case of complete graph communities, every self-loop of the macroscopic configuration model corresponds to a double edge in the hierarchical configuration model. Therefore, equality holds when all communities are complete graphs.

2.3 Clustering coefficient

The clustering coefficient C of a random graph is defined as

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}}. \quad (26)$$

A connected triple is a vertex with edges to two different other vertices. Note that the order of the vertices to which the middle vertex is connected does not matter. The clustering coefficient can be interpreted as the proportion of connected triples that are triangles. In the standard configuration model, the clustering coefficient tends to zero if $\mathbb{E}[D^2] < \infty$ [15]. Thus, in the hierarchical configuration model, we expect that the clustering is entirely caused by triangles inside communities.

Another measure of clustering is the local clustering coefficient for vertices of degree k . This coefficient can be interpreted as the fraction of neighbors of degree k vertices that are connected and is defined as

$$C_k = \frac{\text{number of pairs of connected neighbors of degree } k \text{ vertices}}{k(k-1)/2 \times \text{number of degree } k \text{ vertices}}. \quad (27)$$

As in Section 2.2, let $n_k^{(H)}$ denote the number of vertices in community H with degree equal to k . Furthermore, let $P_v^{(H)}$ denote the number of pairs of neighbors of a vertex $v \in V_H$ within community H that are also neighbors of each other. We denote the clustering coefficient of community H by C_H . Every vertex v in community H has $d_v^{(c)}(d_v^{(c)} - 1)/2$ pairs of neighbors inside H . Hence, the total number of connected triples inside the community is given by $\sum_{v \in V_H} d_v^{(c)}(d_v^{(c)} - 1)/2$. Then, by (26),

$$C_H = \frac{\sum_{v \in V_H} P_v^{(H)}}{\sum_{v \in V_H} d_v^{(c)}(d_v^{(c)} - 1)}. \quad (28)$$

Proposition 8 states that the clustering coefficient of the hierarchical configuration model can be written as a combination of the clustering coefficients inside communities. Let \hat{D} denote the asymptotic degree as in Proposition 5.

Proposition 8. *Let G be a hierarchical configuration model satisfying Conditions 1 and 2, $\lim_{n \rightarrow \infty} \mathbb{E}[D_n^2] = \mathbb{E}[D^2] < \infty$ and $\lim_{N \rightarrow \infty} \mathbb{E}[\hat{D}_N^2] = \mathbb{E}[\hat{D}^2] < \infty$. Then the clustering coefficient $C^{(n)}$ and average clustering coefficient for vertices of degree k , $C_k^{(n)}$, satisfy*

$$C^{(n)} \xrightarrow{\mathbb{P}} C := \frac{2 \sum_H \sum_{v \in V_H} P(H) C_H d_v^{(c)}(d_v^{(c)} - 1)}{\sum_H \sum_{v \in V_H} P(H) d_v(d_v - 1)}, \quad (29)$$

$$C_k^{(n)} \xrightarrow{\mathbb{P}} C_k := \frac{2 \sum_H \sum_{v \in V_H: d_v=k} P(H) P_v^{(H)}}{k(k-1)}. \quad (30)$$

Proof. In the hierarchical configuration model, the number of triples is deterministic. A vertex v with degree d_v has $d_v(d_v - 1)/2$ pairs of neighbors. Thus, the total number of connected triples in G is given by $\sum_H n_H^{(n)} \sum_{v \in V_H} d_v(d_v - 1)/2$, where $n_H^{(n)}$ is the number of type H communities.

Triangles in G can be formed in several ways. First of all, a triangle can be formed by three edges inside the same community. In this case the triangle in G is formed by a triangle in one of its communities H . Another possibility to create a triangle is shown in Figure 1a. The black edges show a community, and the dashed edges are formed by the macroscopic configuration model. This triangle is formed by two intra-community edges, and one edge of the macroscopic configuration model. Figure 2a shows that this inter-community edge is a self-loop of the macroscopic configuration model. One self-loop of the macroscopic configuration model can create multiple triangles; at most $s_i - 2$. Figure 1b shows the case where only one edge of the triangle is an intra-community edge. Figure 2b shows that the two inter-community edges must form a double edge in the macroscopic configuration model. The last possibility is that all three edges of the triangle are inter-community edges as in Figure 1c. This corresponds to a triangle in the macroscopic configuration model (Figure 2c).

Hence, either the triangle was present in H already, or it corresponds to a double edge, self-loop or triangle in $\phi(G)$. Let the number of self-loops, double edges and triangles in $\phi(G)$ be denoted by $W^{(n)}$, $M^{(n)}$ and $T^{(n)}$ respectively. Denote the number of triangles entirely in communities of G by $T_{\text{com}}^{(n)}$. The number of triangles in G is bounded from below by $T_{\text{com}}^{(n)}$. Using (28), we obtain that

$$3T_{\text{com}}^{(n)} = \sum_H \sum_{v \in V_H} n_H^{(n)} P_v^{(H)} = \sum_H \sum_{v \in V_H} n_H C_H d_v^{(c)}(d_v^{(c)} - 1). \quad (31)$$

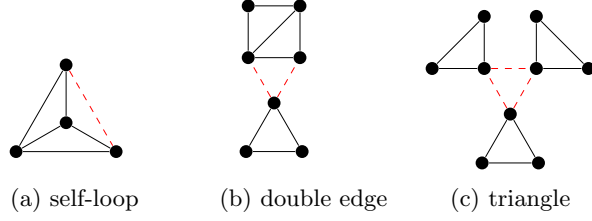


Figure 1: Possibilities to form triangles in the hierarchical configuration model which are not entirely inside communities. Edges between communities (dashed) that add clustering correspond to either a self-loop, a double edge or a triangle in the macroscopic configuration model.

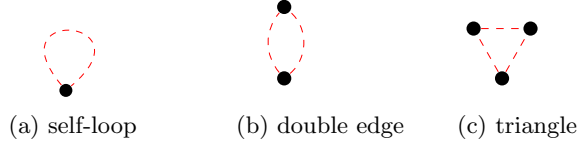


Figure 2: Figure 1 on macroscopic level. The inter-community edges that add clustering correspond to either a self-loop, a double edge or a triangle of the macroscopic configuration model.

Since $P_v^{(H)} \leq d_v^2$, and $\lim_{N \rightarrow \infty} \mathbb{E}[\hat{D}_N] = \mathbb{E}[\hat{D}]$,

$$C^{(n)} = \frac{3 \times \text{number of triangles in } G}{\text{number of connected triples in } G} \geq \frac{3T_{\text{com}}^{(n)}/n}{\sum_H n_H^{(n)} \sum_{v \in V_H} d_v(d_v - 1)/(2n)} \xrightarrow{\mathbb{P}} \frac{2 \sum_H \sum_{v \in V_H} P(H) C_H d_v^{(c)} (d_v^{(c)} - 1)}{\sum_H \sum_{v \in V_H} P(H) d_v(d_v - 1)}. \quad (32)$$

For the upper bound, we use that every self-loop on the community level adds at most $s_i - 2$ triangles, and every triangle and double edge on the community level add at most one triangle. This yields the inequality

$$\text{number of triangles } G \leq T_{\text{com}}^{(n)} + M^{(n)} + S^{(n)} + \sum_{i=1}^{W^{(n)}} (s_{\mathcal{I}_i} - 2). \quad (33)$$

Here the sum is over all communities where a self-loop is present, written as $(\mathcal{I}_i)_{i=1}^{W^{(n)}}$. If a community has multiple self-loops, then the community is counted multiple times in the sum. By [2, Theorem 5]

$$(M^{(n)} + T^{(n)})/n \xrightarrow{\mathbb{P}} 0. \quad (34)$$

in a configuration model with $\mathbb{E}[D^2] < \infty$.

The last term in (33) satisfies

$$\frac{\sum_{i=1}^{W^{(n)}} (s_{\mathcal{I}_i} - 2)}{n} = \frac{W^{(n)} \mathbb{E}[S_n - 2 \mid \text{self-loop}]}{n} \leq \frac{W^{(n)} \max_{i \in [n]} s_i}{n} = \frac{W^{(n)} o(n)}{n} \xrightarrow{\mathbb{P}} 0. \quad (35)$$

The last equality follows because $\mathbb{E}[S] < \infty$, which implies that $\max_i s_i = o(n)$. The convergence follows since the number of self-loops in a configuration model converges to a Poisson distribution with mean ν_D [8, Proposition 7.11], combined with $\mathbb{E}[D^2] < \infty$.

Combining (34) and (35) yields

$$C^{(n)} \leq \frac{3T_{\text{com}}^{(n)} + 3(M^{(n)} + T^{(n)}) + 3 \sum_{i=1}^{W^{(n)}} (s_{\mathcal{I}_i} - 2)}{\sum_H n_H^{(n)} \sum_{v \in V_H} d_v(d_v - 1)/2} \xrightarrow{\mathbb{P}} \frac{2 \sum_H P(H) \sum_{v \in V_H} C_H d_v^{(c)} (d_v^{(c)} - 1)}{\sum_H P(H) \sum_{v \in V_H} d_v(d_v - 1)}. \quad (36)$$

Together with (32) this proves (29).

To prove (30), a similar argument can be used. The number of connected neighbors of vertices of degree k is bounded from below by $\sum_H n_H^{(n)} \sum_{v \in V_H: d_v=k} P_v^{(H)}$, and from above by

$$\sum_H n_H^{(n)} \sum_{v \in V_H: d_v=k} P_v^{(H)} + M^{(n)} + S^{(n)} + \sum_{i=1}^{W^{(n)}} (s_{\mathcal{I}_i} - 2). \quad (37)$$

Then, dividing by $k(k-1)n\hat{p}_k^{(n)}$, where \hat{p}_k is the probability of having a vertex of degree k , and taking the limit yields (30). Note that the assumption that $\mathbb{E}[\hat{D}^2] < \infty$ is not necessary for this clustering coefficient, since $P_v/k(k-1) \leq 1$ for all degree k vertices. \square

3 Percolation

We now consider bond percolation on G , where each edge of G is removed independently with probability $1 - \pi$. We are interested in the critical percolation value and the size of the largest percolating cluster.

Percolating G is the same as first percolating only the edges within communities, and then percolating the edges between communities. For percolation inside a community, only the edges inside a community are removed with probability $1 - \pi$. The half-edges attached to a community are not percolated. Let H_π denote the subgraph of H , where each edge of H has been deleted with probability $1 - \pi$. When percolating a community, it may split into different components. Let $g(H, v, l, \pi)$ denote the probability that the component of H_π containing v has l half-edges. If H_π is still connected, then the component containing v still has d_H outgoing edges for all $v \in V_H$. If H_π is disconnected, then this does not hold. If one of the components of H_π has an outgoing edge, each vertex in another component of H_π cannot reach that edge. Therefore, a vertex in this other component is connected to less than d_H outgoing edges.

To compute the size of the largest percolating cluster, we need the following definitions:

$$p'_k := \frac{\sum_H \sum_{v \in V_H} d_v^{(b)} P(H) g(H, v, k, \pi) / k}{\sum_H \sum_{v \in V_H} \sum_l d_v^{(b)} P(H) g(H, v, l, \pi) / l}, \quad (38)$$

$$h(z) := \sum_{k=1}^{\infty} s p'_k z^{k-1}, \quad (39)$$

$$\lambda := \sum_{k=0}^{\infty} k p'_k. \quad (40)$$

The probabilities $(p'_k)_{k \geq 0}$ can be interpreted as the probability distribution of the outside degrees of the connected parts of communities after percolation inside communities. Then $h(z)$ and λ are the derivative of the probability generating function and the mean of the outside degrees of the components of communities after percolation respectively.

Define D_π^* as the number of outgoing edges after entering a percolated community from a randomly chosen edge. The probability of entering at vertex v in community H , is $P(H) d_v^{(b)} / \mathbb{E}[D]$. After entering H at vertex v , there are in expectation $\sum_{k=1}^{D_H-1} k g(H, v, k+1, \pi)$ edges to other communities (since one edge was used to enter H). Hence,

$$\mathbb{E}[D_\pi^*] = \frac{1}{\mathbb{E}[D]} \sum_H P(H) \sum_{v \in V_H} d_v^{(b)} \sum_{k=1}^{D_H-1} k g(H, v, k+1, \pi). \quad (41)$$

After percolating the inter-community edges, a fraction of π of these edges remain. Thus, after percolating all edges, the expected number of outgoing edges when entering a community is $\pi \mathbb{E}[D_\pi^*]$. We expect the critical value of π to satisfy $\pi \mathbb{E}[D_\pi^*] = 1$, i.e., the expected number of edges to other communities is one, after entering a community from a randomly chosen edge. The next theorem states that this is indeed the critical percolation value:

Theorem 9. Assume G is a hierarchical configuration model satisfying Conditions 1 and 2. The critical value of the percolation parameter, π_c of G satisfies

$$\pi_c = \frac{1}{\mathbb{E}[D_{\pi_c}^*]}. \quad (42)$$

Furthermore,

- (i) For $\pi > \pi_c$, the size of the largest component of the percolated graph satisfies

$$\frac{v(\mathcal{C}_{\max})}{N} \xrightarrow{\mathbb{P}} \frac{1}{\mathbb{E}[S]} \sum_{l=1}^{\infty} \sum_H \sum_{v \in V_H} P(H) g(H, v, l, \pi) (1 - (1 - \sqrt{\pi} + \sqrt{\pi}\xi)^l) > 0, \quad (43)$$

where ξ is the unique solution in $(0, 1)$ of

$$\sqrt{\pi}h(1 - \sqrt{\pi} + \sqrt{\pi}\xi) + (1 - \sqrt{\pi})\lambda = \lambda\xi. \quad (44)$$

- (ii) For $\pi \leq \pi_c$, $v(\mathcal{C}_{\max})/N \xrightarrow{\mathbb{P}} 0$.

Note that for the standard configuration model, (42) simplifies to $\pi_c = \mathbb{E}[D]/\mathbb{E}[D(D-1)]$, since in that case, for a vertex of degree d , $g(v, v, k, \pi) = \mathbb{1}_{\{k=d\}}$. Furthermore, $\pi_c = 0$ when for any $\pi > 0$, the expected number of edges to other communities is infinite when entering a community via a uniformly chosen edge.

Proof. The proof of Theorem 9 has a similar structure as the proof of Theorem 1 in [4]. The proof consists of three key steps:

- (i) First, each edge within each community is removed with probability $1 - \pi$. This may split the community into several connected components. We find the distribution of the outside degrees of the connected components of the percolated communities, which is given by p'_k , as in (38). We identify vertices that are in the same connected component of a community. Lemma 10 shows that this results in a graph $\phi(G_\pi)$ that is distributed as a configuration model with degree probabilities p'_k .
- (ii) We then remove each edge between communities with probability $1 - \pi$. Results of [11] can now be applied on the configuration model with distribution p'_k to find the critical percolation value and the size of the giant hierarchical component.
- (iii) Next, we translate the number of communities in the largest percolated hierarchical component to the number of vertices. Then we show that this is indeed the largest component of the percolated graph.

Auxiliary graph. We introduce the auxiliary graph $\phi(\bar{G})$, defined for every subgraph $\bar{G} \subset G$, and obtained by identifying the vertices that belonged to the same community in G , and are connected in \bar{G} [4]. Hence, in $\phi(\bar{G})$ every vertex represents a connected part of a community. Figure 3 illustrates $\phi(\bar{G})$. For a hierarchical configuration model G , the graph $\phi(G)$ is a configuration model where communities of G are collapsed into single vertices.

Lemma 10. Let G be a hierarchical configuration model satisfying Conditions 1 and 2. Let G_π denote the subgraph of G where each edge inside each community is removed with probability $1 - \pi$. Then the graph $\phi(G_\pi)$ is distributed as a configuration model with degree probabilities p'_k given in (38).

Proof. We independently delete each edge within each community with probability $1 - \pi$. We want to find the degree distribution of $\phi(G_\pi)$. Let $M^{(n)}(H, v, k, \pi)$ denote the number of connected components of the percolated versions of community H containing vertex v and having outside degree k . Each community of shape H has an equal probability that the component containing

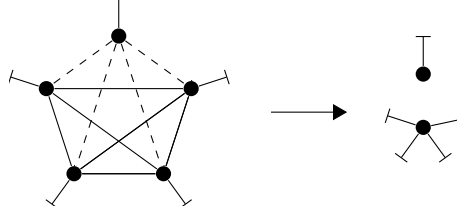


Figure 3: Left, a subgraph \bar{H} of a community H , dashed lines are not in \bar{H} , but are present in H , solid lines are present in \bar{H} . The graph $\phi(\bar{H})$ is shown on the right.

v has outside degree k given by $g(H, v, k, \pi)$. Furthermore, the probability that a randomly chosen community has shape H is independent of the probability that after percolation there are k outgoing half-edges in a community of shape H . Therefore, given the number of type H communities $n_H^{(n)}$, $M^{(n)}(H, v, k, \pi) \sim \text{Bin}(n_H^{(n)}, g(H, v, k, \pi))$. Thus, by the weak law of large numbers,

$$\frac{M^{(n)}(H, v, k, \pi)}{n} = \frac{M^{(n)}(H, v, k, \pi)}{n_H^{(n)}} \frac{n_H^{(n)}}{n} \xrightarrow{\mathbb{P}} P(H)g(H, v, k, \pi). \quad (45)$$

Let $N^{(n)}(H, k, \pi)$ denote the total number of connected components of the percolated versions of H having outside degree k . This number can be obtained by counting the number of half-edges of all connected components of percolated graphs with outside degree k , and then dividing by k . Each vertex v in such a percolated community contributes $d_v^{(b)}$ to the number of half-edges of the percolated community. Thus,

$$N^{(n)}(H, k, \pi) = \sum_{v \in V_H} d_v^{(b)} M^{(n)}(H, v, k, \pi) / k. \quad (46)$$

Let \tilde{n} denote the number of vertices in $\phi(G_\pi)$, so that $\tilde{n} = \sum_H \sum_k N^{(n)}(H, k, \pi)$. Similarly, the number vertices of degree k in $\phi(G_\pi)$ is denoted by $\tilde{n}_k = \sum_H N^{(n)}(H, k, \pi)$. Furthermore, $\sum_k N^{(n)}(H, k, \pi) / n \leq P_n(H) s_H$, and therefore by Condition 1, (45) and (46),

$$\tilde{n} / n \xrightarrow{\mathbb{P}} \sum_H \sum_{k=1}^{D_H} \sum_{v \in V_H} d_v^{(b)} P(H) g(H, v, k, \pi) / k, \quad (47)$$

Then also

$$\frac{N^{(n)}(H, k, \pi)}{\tilde{n}} = \frac{N^{(n)}(H, k, \pi) / n}{\tilde{n} / n} \xrightarrow{\mathbb{P}} \frac{\sum_{v \in V_H} d_v^{(b)} P(H) g(H, v, k, \pi) / k}{\sum_H \sum_{v \in V_H} \sum_l d_v^{(b)} P(H) g(H, v, l, \pi) / l}. \quad (48)$$

Hence, the proportion of vertices in $\phi(G_\pi)$ with degree k goes to

$$\frac{\tilde{n}_k}{\tilde{n}} = \sum_H \frac{N^{(n)}(H, k, \pi)}{\tilde{n}} \xrightarrow{\mathbb{P}} p'_k. \quad (49)$$

Since the edges between communities in G were paired at random, this means that the graph $\phi(G_\pi)$ is distributed as a configuration model with degree probabilities p'_k . \square

Using Lemma 10, we now prove Theorem 9.

Step (i). Lemma 10 proves that $\phi(G_\pi)$ is distributed as a configuration model with degree probabilities p'_k .

Step (ii). $\phi(G_\pi)$ and $\phi(G)$ have $\sum_k k \tilde{n}_k$ and $\sum_k k n_k$ half-edges, respectively. Since only edges inside communities have been deleted, $\sum_k k n_k$ equals $\sum_k k \tilde{n}_k$. By Condition 2(i), $\sum_k k n_k / n \rightarrow$

$\mathbb{E}[D]$. Furthermore, by (47) \tilde{n}/n converges, hence $\sum_k k\tilde{n}_k/\tilde{n}$ converges. Therefore we can apply Theorem 3.9 from [11], which states that after percolation, a configuration model with degree probabilities p'_k has a giant component if

$$\pi \sum_k k(k-1)p'_k > \sum_k kp'_k. \quad (50)$$

From Theorem 3 we know that a giant hierarchical component is also a giant component in G , and a hierarchical component of size $o_{\mathbb{P}}(n)$ is a component of size $o_{\mathbb{P}}(N)$. Hence, the giant component emerges precisely when the giant hierarchical component emerges. Substituting (49) gives for the critical percolation value π_c that,

$$\begin{aligned} \pi_c &= \frac{\sum_k kp'_k}{\sum_k k(k-1)p'_k} = \frac{\sum_H \sum_{v \in V_H} d_v^{(b)} \sum_{k=1}^{D_H} P(H)g(H, v, k, \pi_c)k/k}{\sum_H \sum_{v \in V_H} d_v^{(b)} \sum_{k=1}^{D_H} P(H)g(H, v, k, \pi_c)k(k-1)/k} \\ &= \frac{\sum_H \sum_{v \in V_H} d_v^{(b)} P(H)}{\sum_H \sum_{v \in V_H} d_v^{(b)} \sum_{k=1}^{D_H} P(H)g(H, v, k, \pi_c)(k-1)} \\ &= \frac{\mathbb{E}[D]}{\sum_H \sum_{v \in V_H} \sum_{k=1}^{D_H-1} d_v^{(b)} P(H)g(H, v, k+1, \pi_c)k} \\ &= \frac{1}{\mathbb{E}[D_{\pi_c}^*]}. \end{aligned} \quad (51)$$

Step (iii). Now assume that $\pi > \pi_c$. The number of degree r vertices in the largest component of $\phi(G_\pi)$ satisfies $v_r(\mathcal{C}_{\max}^H)/\tilde{n} \xrightarrow{\mathbb{P}} \sum_{l \geq r} b_{lr}(\sqrt{\pi})p'_l(1-\xi^r)$ [11], with $b_{lr}(\sqrt{\pi}) = \binom{l}{r}\sqrt{\pi}^r(1-\sqrt{\pi})^{l-r}$ as the probability that a binomial with parameters l and $\sqrt{\pi}$ takes value r , and ξ is as in (44). To translate the number of percolated communities in the largest component into the number of vertices in the largest component, we want to know the expected number of vertices in the largest component that are in a percolated community with outside degree k . The size of a percolated community is independent of being in the largest hierarchical component, but does depend on the outside degree of the percolated community. The total number of vertices in connected percolated components with outside degree k is given by $\sum_H \sum_{v \in V_H} M^{(n)}(H, v, k, \pi)$, and the total number of percolated communities of outside degree k is given by $\sum_H N^{(n)}(H, k, \pi)$. Furthermore, $\sum_{v \in V_H} M^{(n)}(H, v, k, \pi)/n \leq P_n(H)s_H$. Hence, by Condition 2, the expected size of a percolated community, given it has outside degree k , satisfies

$$\begin{aligned} \mathbb{E}[S_\pi \mid \text{outside degree } k] &= \frac{\sum_H \sum_{v \in V_H} M^{(n)}(H, v, k, \pi)}{\sum_H N^{(n)}(H, k, \pi)} \\ &= \frac{\sum_H \sum_{v \in V_H} M^{(n)}(H, v, k, \pi)/n}{\sum_H N^{(n)}(H, k, \pi)/n} \\ &\xrightarrow{\mathbb{P}} \frac{\sum_H \sum_{v \in V_H} P(H)g(H, v, k, \pi)}{\sum_H \sum_{v \in V_H} d_v^{(b)} P(H)g(H, v, k, \pi)/k}. \end{aligned} \quad (52)$$

Since $v_r(\mathcal{C}_{\max})/N \leq \hat{p}_r^{(n)}$, which sums to one, we can write for the number of vertices in the largest component of $\phi(G_\pi)$

$$\begin{aligned} \frac{v(\mathcal{C}_{\max})}{N} &= \sum_{r=0}^{\infty} \frac{v_r(\mathcal{C}_{\max})/\tilde{n}}{N/n \cdot n/\tilde{n}} \xrightarrow{\mathbb{P}} \sum_{r=0}^{\infty} \frac{\sum_{l \geq r} b_{lr}(\sqrt{\pi})p'_l(1-\xi^r)\mathbb{E}[S_\pi \mid \text{outside degree } l]}{\mathbb{E}[S]/\sum_H \sum_{v \in V_H} \sum_k d_v^{(b)} P(H)g(H, v, k, \pi)/k} \\ &= \frac{\sum_{r=0}^{\infty} \sum_{l \geq r} b_{lr}(\sqrt{\pi})(1-\xi^r)p'_l}{\mathbb{E}[S]/\sum_H \sum_{v \in V_H} \sum_k d_v^{(b)} P(H)g(H, v, k, \pi)/k} \frac{\sum_H \sum_{v \in V_H} P(H)g(H, v, l, \pi)}{\sum_H \sum_{v \in V_H} d_v^{(b)} P(H)g(H, v, l, \pi)/l} \\ &= \frac{\sum_{l=0}^{\infty} \sum_{r=0}^l b_{lr}(\sqrt{\pi})(1-\xi^r)\frac{p'_l}{p'_l} \sum_H \sum_{v \in V_H} P(H)g(H, v, l, \pi)}{\mathbb{E}[S]} \\ &= \frac{\sum_{l=0}^{\infty} (1 - (1 - \sqrt{\pi} + \sqrt{\pi}\xi)^l) \sum_H \sum_{v \in V_H} P(H)g(H, v, l, \pi)}{\mathbb{E}[S]}. \end{aligned} \quad (53)$$

Any other component of $\phi(G_\pi)$ has w.h.p. size of $o_{\mathbb{P}}(\tilde{n})$ by [11, Theorem 3.9]. As shown in the proof of Theorem 3, any component of size $o_{\mathbb{P}}(\tilde{n})$ in $\phi(G_\pi)$ is a component of $o_{\mathbb{P}}(N)$ in the total graph. Hence, w.h.p. \mathcal{C}_{\max} is the largest component of the percolated graph G_π .

When $\pi < \pi_c$, the largest component of $\phi(G_\pi)$ satisfies $v(\mathcal{C}_{\max}^H)/\tilde{n} \xrightarrow{\mathbb{P}} 0$ [11]. Again, by the analysis of Theorem 3, this component is of size $o_{\mathbb{P}}(N)$ in the original graph. \square

Equation (44) also has an intuitive explanation. Let Q be the distribution of the community outside degrees after percolation when following a randomly chosen half-edge. Then we can interpret ξ as the extinction probability of a branching process with offspring distribution Q . Percolating the inter-community edges with probability $1 - \pi$ is the same as deleting each half-edge with probability $1 - \sqrt{\pi}$. Then, with probability $1 - \sqrt{\pi}$ the randomly chosen half-edge is paired to a deleted half-edge, in which case the branching process goes extinct. With probability $\sqrt{\pi}$, the half-edge leads to a half-edge which still exists after percolation, and leads to a community. The probability generating function of the number of half-edges pointing out of this community before percolating the half-edges is $\frac{1}{\lambda}h(\xi)$. Since the number of half-edges after percolation is binomial given the number of half-edges that were present before percolation, the probability generating function of the number of half-edges pointing out of a community entered by a randomly chosen half-edge is $\frac{1}{\lambda}h(1 - \sqrt{\pi} + \sqrt{\pi}\xi)$. Combining this yields (44).

The case $\mathbb{E}[D^2] = \infty$. In the standard configuration model $\pi_c = 0$ precisely when $\mathbb{E}[D^2] = \infty$. In the hierarchical configuration model, this may not be true, since it is possible to construct communities with large outside degrees, while all individual vertices have a low degree. An example of such a community structure is the hierarchical configuration model, where each community is a line graph H_l of l vertices with probability \bar{p}_l , where each vertex has outside degree one. Figure 4a illustrates H_l for $l = 5$. We assume that \bar{p}_l obeys the power-law $\bar{p}_l = cl^{-\alpha}$, with $\alpha \in (2, 3)$. Then $\mathbb{E}[D] < \infty$, but $\mathbb{E}[D^2] = \infty$. Hence, communities may have large outside degrees. However, G is a 3-regular graph, so no individual vertex has high degree. From this fact, we can already conclude that $\pi_c \neq 0$. Suppose $\pi_c < \frac{1}{2}$. Then, after percolation every vertex has less than two expected neighbors. Hence, there is no giant component w.h.p. We can also use Theorem 9 to show that $\pi_c \neq 0$. We compute the denominator of (42), and show that it is finite. We have

$$\sum_{v \in V_H} d_v^{(b)} g(H_l, v, k, \pi_c) = \begin{cases} 2k\pi_c^{k-1}(1 - \pi_c) + k\pi_c^{k-1}(1 - \pi_c)^2(l - k - 1) & \text{if } k < l, \\ k\pi_c^{k-1} & \text{if } k = l. \end{cases} \quad (54)$$

This gives

$$\sum_{k=1}^{l-1} k \sum_{v \in V_H} d_v^{(b)} g(H_l, v, k+1, \pi_c) = \frac{2\pi_c(\pi_c^l + l(1 - \pi_c)) - 1}{(1 - \pi_c)^2}. \quad (55)$$

Using that $\bar{p}_l = cl^{-\alpha}$ gives for (41)

$$\begin{aligned} \mathbb{E}[D_{\pi_c}^*] &= \frac{1}{\mathbb{E}[D]} \sum_H P(H) \sum_{v \in V_H} d_v^{(b)} \sum_{k=1}^{D_H-1} kg(H, v, k+1, \pi_c) \\ &= \frac{1}{\mathbb{E}[D]} \sum_{l=1}^{\infty} cl^{-\alpha} \frac{2\pi_c(\pi_c^l + l(1 - \pi_c)) - 1}{(1 - \pi_c)^2} \\ &= \frac{1}{\mathbb{E}[D]} \frac{2\pi_c}{(1 - \pi_c)^2} \left(-1 + (1 - \pi_c)\mathbb{E}[D] + \sum_{l=1}^{\infty} c\pi_c^l l^{-\alpha} \right). \end{aligned} \quad (56)$$

From (56) we see that $\pi_c = 0$ is not a solution of (42). Hence, $\pi_c \neq 0$, even though $\mathbb{E}[D^2] = \infty$.

Infinite second moment of degree. When the second moment of the degree distribution as defined in Proposition 5 is infinite, π_c also does not have to be zero. It is possible to ‘hide’ all

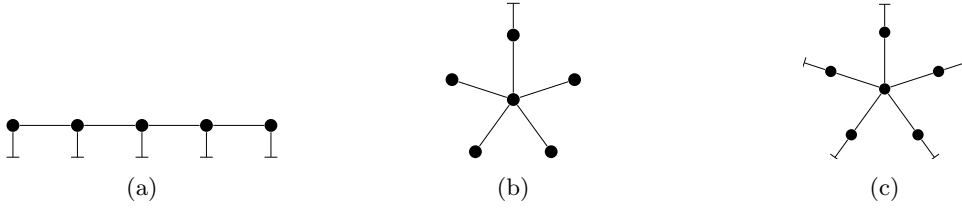


Figure 4: Communities with $l = 5$.

vertices of high degree inside communities that have small outside degrees. The small outside degrees of the communities make it difficult to leave the community in percolation. One example of such a community structure is the case in which each community is a star-shaped graph with l endpoints with probability p_l . One vertex in the graph has outside degree one, and all the other vertices have outside degree zero. Figure 4b illustrates the star-shaped graph for $l = 5$. Since each community has only one outgoing edge, there cannot be a giant component in G . We can also see this from Theorem 3, since $\mathbb{E}[D] = \mathbb{E}[D^2] = 1$. By Proposition 5, the degree distribution equals

$$\hat{p}_k = \begin{cases} (\sum_{l=1}^{\infty} (l-1)p_l + p_1) / \mathbb{E}[S] & \text{if } k = 1, \\ (1 + p_2) \mathbb{E}[S] & \text{if } k = 2, \\ p_k / \mathbb{E}[S] & \text{if } k > 2. \end{cases} \quad (57)$$

When p_l is a probability distribution with infinite second moment, the second moment of \hat{p}_l is also infinite. Hence, the degree distribution of the hierarchical configuration model G has infinite second moment, while there is no giant component, so that certainly $\pi_c \neq 0$.

A sufficient condition for $\pi_c = 0$. By (42),

$$\begin{aligned} \pi_c &= \frac{\mathbb{E}[D]}{\sum_G P(H) \sum_{v \in V_H} d_v^{(b)} \sum_{k=1}^{D_H-1} kg(H, v, k+1, \pi_c)} \\ &\leq \frac{\mathbb{E}[D]}{\sum_G P(H) \sum_{v \in V_H} d_v^{(b)2}} \end{aligned} \quad (58)$$

Hence, $\sum_H \sum_{v \in V_H} P(H) d_v^{(b)2} = \infty$ is a sufficient condition for $\pi_c = 0$. This condition can be interpreted as an infinite second moment of the outside degrees of individual vertices. However, it is not a necessary condition. It is possible to construct a community where all individual vertices have a small outside degree, but are connected to a vertex with high degree. Consider for example the star community of Figure 4c, with one vertex in the middle, linked to l other vertices. The l other vertices have outside degree one, and the middle vertex has outside degree zero, hence all vertices have a small outside degree. However, the middle vertex can have a high degree. Let each community be a star-shaped community with l outgoing edges with probability \bar{p}_l . We can calculate that $\pi_c = \sum_l \bar{p}_l l(l-1)\pi_c^2$. Hence, if we choose \bar{p}_l with finite first moment and infinite second moment, $\pi_c = 0$. However, $\sum_H \sum_{v \in V_H} d_v^{(b)2} = \sum_l l \bar{p}_l = \mathbb{E}[D] < \infty$.

4 Existing graph models with a community structure

In this section, we show how three existing random graph models with community structure fit within the hierarchical configuration model.

4.1 Trapman's household model

Trapman [18] replaces vertices in a configuration model by households in the form of complete graphs, such that the degree distribution of the resulting graph is p_k . To achieve this, each

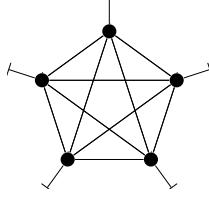


Figure 5: A household of size 5

community is a single vertex of degree k with probability $(1 - \gamma)p_k$, or a complete graph of size k with probability $\gamma\bar{p}_k$. Here \bar{p}_k , the probability that a certain clique has degree k , is given by

$$\bar{p}_k = k^{-1}p_k\mathbb{E}[W^{-1}]^{-1}, \quad (59)$$

where W is a random variable satisfying $\mathbb{P}(W = k) = p_k$. Each vertex of the complete graph has one edge to another community. Figure 5 illustrates a household of size 5. This model is a special case of the hierarchical configuration model with

$$H_i = \begin{cases} (K_k, (1, \dots, 1)) & \text{w.p. } \gamma\bar{p}_k, \\ (v, (k)) & \text{w.p. } (1 - \gamma)p_k, \end{cases} \quad (60)$$

where K_k is a complete graph on k vertices.

We now check when (60) satisfies Conditions 1 and 2. The assumption $\mathbb{P}(D = 2) < 1$ is satisfied if and only if $p_2 < 1$. The expected outside degree of a community is given by

$$\mathbb{E}[D] = (1 - \gamma) \sum_k kp_k + \gamma \sum_k k\bar{p}_k = (1 - \gamma)\mathbb{E}[W] + \frac{\gamma}{\mathbb{E}[W^{-1}]}. \quad (61)$$

Hence, $\mathbb{E}[D] < \infty$ if $\mathbb{E}[W] < \infty$ and $\mathbb{E}[W^{-1}] \neq 0$. By Jensen's inequality, $\mathbb{E}[W^{-1}] \geq \mathbb{E}[W]^{-1} > 0$, hence $\mathbb{E}[D] < \infty$ if and only if $\mathbb{E}[W] < \infty$. For every community in this model, its size is smaller than or equal to its outside degree, so that also $\mathbb{E}[S] < \infty$ if $\mathbb{E}[D] < \infty$. Thus, Conditions 1 and 2 hold if $\mathbb{E}[W] < \infty$ and $p_2 < 1$. Under these conditions we can apply the results for the hierarchical configuration model as derived in Sections 2 and 3.

Suppose p_k follows a power-law with exponent α . Then by (59) \bar{p}_k follows a power-law with exponent $\alpha - 1$, and the distribution of the outside degrees D is a mixture of a power-law with exponent α and a power-law with exponent $\alpha - 1$. Thus, the power-law shift described in Section 2.2.1 does not occur, since in this household model, not all communities are dense. For Trapman's household model, the power-law shift only occurs if $\gamma = 1$, in which case all communities are households.

4.2 Lelarge and Coupechoux' household model

Another model that takes complete graphs as communities is the model of Coupechoux and Lelarge [4]. This model is very similar to Trapman's model. Again, each community is either a complete graph or a single vertex. In contrast to [18], the probability that a certain community is a clique is dependent on the degree of the clique. Each vertex of degree k in the macroscopic configuration model is replaced by a complete graph with probability γ_k . This graph can be modeled as a hierarchical configuration model with

$$H_i = \begin{cases} (K_k, (1, \dots, 1)) & \text{w.p. } \gamma_k\bar{p}_k \\ (v, (k)) & \text{w.p. } (1 - \gamma_k)\bar{p}_k, \end{cases} \quad (62)$$

where $(\bar{p}_k)_{k \geq 1}$ is a probability distribution. Since the outside degrees of all communities have distribution $\mathbb{P}(D = k) = \bar{p}_k$, Condition 2 holds if the probability distribution \bar{p}_k has finite mean

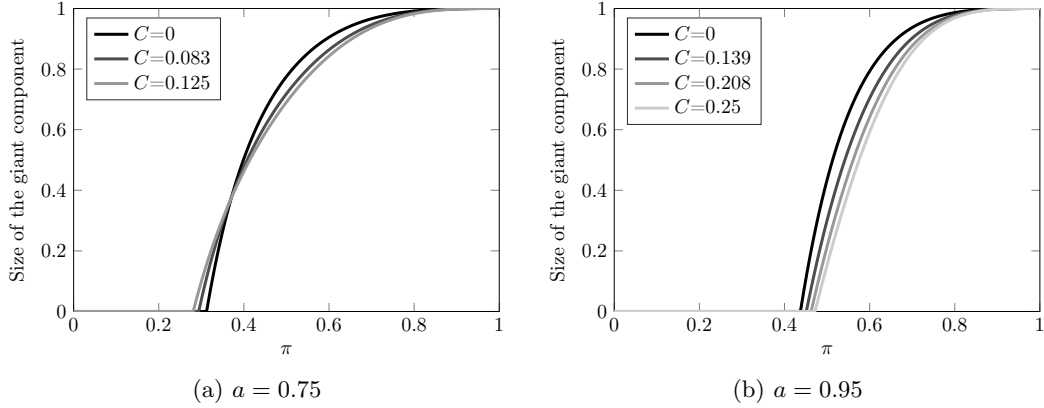


Figure 6: Size of giant component after bond percolation with probability π in a household model with $p_3 = a$ and $p_6 = 1 - a$ for various clustering coefficients C . If $a = 0.75$, adding clustering decreases the critical percolation value, whereas if $a = 0.95$, adding clustering increases the critical percolation value.

and $\bar{p}_2 < 1$. The size of a community is always smaller or equal than its outside degree, so that also $\mathbb{E}[S] < \infty$ if \bar{p}_k has finite mean. Thus, Conditions 1 and 2 hold if \bar{p}_k has finite mean and $\bar{p}_2 < 1$.

If these conditions on \bar{p}_k hold, then the degree distribution p_k of the resulting graph can be obtained from Proposition 5,

$$p_k = \frac{(k\gamma_k + (1 - \gamma_k))\bar{p}_k}{\sum_{i \geq 0} (i\gamma_i + (1 - \gamma_i))\bar{p}_i}. \quad (63)$$

Suppose that $\gamma_k \geq \gamma > 0$. Then, in contrast to Trapman's household model in Subsection 4.1, the degree distribution of the edges between communities, \bar{p}_k , follows a power-law with exponent $\alpha + 1$ if the degree distribution p_k follows a power-law with exponent α .

As an example of such a household model, consider a graph with $p_3 = a$ and $p_6 = 1 - a$ and a tunable clustering coefficient. We take $\gamma_6 = 0$, but increase γ_3 , while the degree distribution remains the same. Thus, the graph consists of only single vertices of degree 6, single vertices of degree 3 and triangle communities. Since we increase γ_3 , the number of triangles increases, so that also the clustering coefficient increases. Figures 6a and 6b show the size of the giant component under percolation for different values of the clustering coefficient using $a = 0.75$ and $a = 0.95$ respectively. In the case where $a = 0.75$, clustering decreases the value of π_c , whereas if $a = 0.95$, clustering increases the value of π_c . This illustrates that the influence of clustering on bond percolation of a random graph is non-trivial. In two similar random graph models, introducing clustering has a different effect.

4.3 Configuration model with triangles

A third random graph model with clustering is the model by Newman [14]. In this model, each vertex v has an edge-degree $d_v^{(1)}$ and a triangle degree $d_v^{(2)}$, denoting the number of triangles that the vertex is part of. Then a random graph is formed by pairing edges at random and pairing triangles at random. Even though this model does not explicitly replace vertices in a configuration model by communities, it is also a special case of the hierarchical configuration model if some conditions on the degrees are satisfied. The communities in this model are the connected components consisting only of triangles. Figure 7 shows two possible realizations of such communities.

From results derived in [14], we can find the probability generating function $h_r(z)$ of the number of vertices in triangles that can be reached from a uniformly chosen triangle, and the probability generating function $h_{S^*}(z)$ of the size of the triangle component of a randomly chosen vertex, that

together satisfy

$$h_r(z) = zg_q(h_r^2(z)), \quad h_{S^*}(z) = zg_p(h_r^2(z)), \quad (64)$$

where g_q is the probability generating function of the size-biased distribution of the triangle degrees, and g_p the probability generating function of the triangle degree distribution. In the hierarchical configuration model, $h_{S^*}(z)$ can be interpreted as the probability generating function of the size-biased community sizes. Thus, the mean size-biased community size is given by

$$\mathbb{E}[S^*] = 1 + \frac{2\mathbb{E}[D^{(2)}]}{3 - 2\mathbb{E}[D^{(2)*}]}, \quad (65)$$

where $D^{(2)*}$ is the size-biased distribution of the triangle degrees. Since $\mathbb{E}[S^*] \geq \mathbb{E}[S]$, Condition 1(ii) is satisfied if $\mathbb{E}[D^{(2)*}] < \frac{3}{2}$.

The mean outside degree of a community is given by

$$\mathbb{E}[D] = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \sum_{v \in G_i} d_v^{(1)}}{n} = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^N d_i^{(1)}/N}{n/N} = \mathbb{E}[S]\mathbb{E}[D^{(1)}]. \quad (66)$$

Hence, Conditions 1 and 2 are satisfied if $\mathbb{E}[D^{(2)*}] < \frac{3}{2}$ and $\mathbb{E}[D^{(1)}] < \infty$. When these conditions are satisfied, the condition for emergence of a giant component is

$$\frac{\mathbb{E}[D^{(1)2}]\mathbb{E}[S] - \mathbb{E}[D^{(1)}]\mathbb{E}[S]^2}{\mathbb{E}[D^{(1)}]\mathbb{E}[S]} = \frac{\mathbb{E}[D^{(1)2}] - \mathbb{E}[D^{(1)}]^2}{\mathbb{E}[D^{(1)}]} > 1. \quad (67)$$

Therefore, as long as $\mathbb{E}[D^{(2)*}] < \frac{3}{2}$, the emergence of the giant component only depends on the edge degree distribution.

To apply the results of the hierarchical configuration model, we need the probability $P(H)$ that a randomly chosen community is of type H . This probability is not easy to obtain, but it can be approximated using a branching process. The branching process starts at a vertex, and explores the component of triangles. The first generation of the branching process has $Z_0 = 1$. The first offspring, Z_1 is distributed as $2D^{(2)}$. All other offspring, Z_i for $i > 1$ is distributed as $\sum_{j=1}^{Z_{i-1}} 2(D_j^{(2)*} - 1)$. Here $D_j^{(2)*}$ are independent copies, distributed as $D^{(2)*}$. In this branching process approximation, cycles of triangles are ignored. The size-biased probability of having a specific community H can be obtained by summing the probabilities of the possible realizations of the branching process when exploring graph H . This probability is size-biased, since when starting at an arbitrary vertex, the probability of starting in a larger community is higher. This probability then needs to be transformed to the probability of obtaining graph H .

To compute the size of the giant component after percolation from (43), $g(H, v, k, \pi)$ is needed for every community shape H . This is difficult to obtain, since it largely depends on the shape of the community, and there are infinitely many possible community shapes. Figure 7 shows an example of why the shape of a community matters. When percolating the left community, the probability that the red vertex is connected to k other vertices is smaller than for the graph on the right. For this reason, we approximate (43) numerically using the branching process described above. In [14], Newman gives expressions for the size of the largest percolating cluster. Figure 8 compares the size of the giant component computed in that way with a numerical approximation of (43). We see that indeed the equations from [14] give the same results for the largest percolating cluster as (43).

5 Stylized networks

In this section, we study two stylized examples of community structures. The first example gives a community type that decreases the critical percolation value compared with a configuration model with the same degree distribution. The second example increases the critical percolation value when compared to a configuration model.

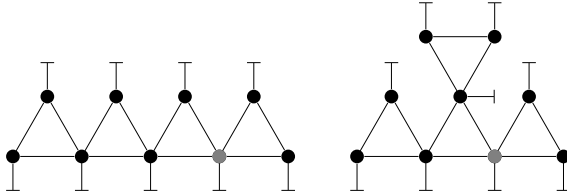


Figure 7: Two possible communities with 4 triangles. In the left community, 6 other nodes can be reached from the gray node within 2 steps, in the right community 8 nodes

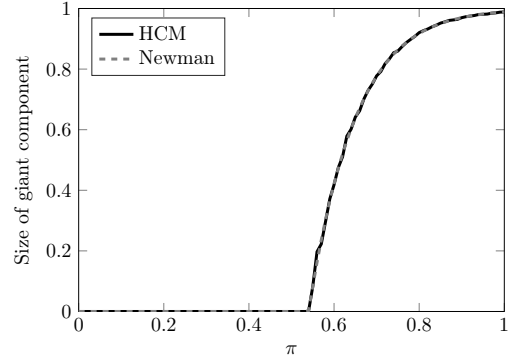


Figure 8: The size of the giant percolating cluster calculated by (43) (hierarchical configuration model) and from results in [14] (Newman) agree.

5.1 A community structure that decreases π_c

As an example of a community structure that decreases π_c , we consider a hierarchical configuration model where with probability ϕ a community is given by H_1 : a path of l vertices, with a half-edge at each end of the path as illustrated in Figure 9. With probability $1 - \phi$ the community is H_2 : a vertex with three half-edges. The degree distribution of this hierarchical configuration model can be found using Proposition 5 and is given by

$$p_k = \begin{cases} \frac{l\phi}{l\phi+1-\phi} & \text{if } k = 2, \\ \frac{(1-\phi)}{l\phi+1-\phi} & \text{if } k = 3, \\ 0 & \text{otherwise.} \end{cases} \quad (68)$$

We have $\mathbb{E}[D] = 2\phi + 3(1 - \phi)$. Furthermore, $g(H_1, v, 2, \pi) = \pi^{l-1}$ for all $v \in H_1$. In H_2 there is no percolation inside the community, hence $g(H_2, v, 3, \pi) = 1$. Equation (42) now gives:

$$\pi_c = \frac{3 - \phi}{2\phi\pi_c^{l-1} + 6(1 - \phi)}, \quad (69)$$

hence $2\phi\pi_c^l + 6(1 - \phi)\pi_c - 3 + \phi = 0$.

Now we let the degree distribution as defined in (68) remain the same, while changing the length of the path communities l . If in the total graph, we want to have a fraction of a vertices of degree 3, then $a = p_3 = \frac{1-\phi}{1-\phi+l\phi}$. Hence, $\phi = \frac{1-a}{1-a+la}$. In this way, we obtain hierarchical configuration models with the same degree distribution, but with different values of l . Figure 10 shows the size of the largest component as calculated by (43) for $a = 1/3$. As l increases, π_c decreases. Hence, adding this community structure ‘helps’ the diffusion process. This can be explained by the fact that increasing l decreases the number of line communities. Therefore, more vertex communities will be connected to one another, which decreases the value of π_c . Another interesting observation is that the size of the giant component is non-convex in π . These non-convex shapes can be explained intuitively. As the lines get longer, there are fewer and fewer of them, since the degree distribution remains the same. Hence, if l is large, there will only be a few long lines. These lines have $\pi_c \approx 1$. Since there are only a few lines, almost all vertices of degree 3 will be paired to one another. The critical value for percolation on a configuration model with only vertices of degree 3 is 0.5. Hence, for this hierarchical configuration model with l large we will see the vertices of degree 3 appearing in the giant component as $\pi = 0.5$, and the vertices in the lines as $\pi = 1$.



Figure 9: A line community with $l = 5$

5.2 A community structure that increases π_c

As an example of a community structure that inhibits the diffusion process, consider a configuration model with intermediate vertices [10]: a configuration model where every edge is replaced by two edges with a vertex in between them. This is equal to a hierarchical configuration model with star-shaped communities as in Figure 4c: one vertex that is connected to l other vertices. Each of the l other vertices has outside degree one. The vertex in the middle is not connected to other communities. We consider a hierarchical configuration model where all communities are stars of the same size. Therefore all star-shaped communities have the same number of outgoing edges, and $\mathbb{E}[D] = l$.

The degree distribution of this hierarchical configuration model is given by

$$p_k = \begin{cases} \frac{l}{l+1} & \text{if } k = 2, \\ \frac{1}{l+1} & \text{if } k = l, \\ 0 & \text{otherwise.} \end{cases} \quad (70)$$

Under percolation, the connected component of a vertex v at the end point of a star can link to other half-edges only if the edge to the middle vertex is present. Then the number of half-edges to which v is connected, is binomially distributed, so that $g(H, v, k, \pi) = \pi \binom{l-1}{k-1} \pi^{k-1} (1-\pi)^{l-k}$ for $k \geq 2$. Then (41) gives

$$\begin{aligned} \mathbb{E}[D_\pi^*] &= \frac{1}{\mathbb{E}[D]} \sum_H \sum_k \sum_{v \in V_H} P(H) d_v^{(b)} k g(H, v, k+1, \pi) \\ &= \pi \sum_{k \geq 1} k \pi^k (1-\pi)^{l-k-1} \binom{l-1}{k} = (l-1) \pi^2. \end{aligned} \quad (71)$$

Then equation (42) yields $\pi_c = (l-1)^{-1/3}$.

Now we consider a configuration model with the same degree distribution (70). For this configuration model, $\pi_c = \frac{3l}{4l+l^2-3l} = \frac{3}{l+1}$. Figure 11 shows the size of the giant component of the hierarchical configuration model compared with a configuration model with the same degree distribution for different values of l . This hierarchical configuration has a higher critical percolation value than its corresponding configuration model. Intuitively, this can be explained from the fact that all vertices with a high degree are ‘hidden’ behind vertices of degree 2, whereas in the configuration model, vertices of degree l may be connected to one another.

Combined with the previous example, we see that adding communities may lead to a higher critical percolation value or a lower one. Furthermore, the size of the giant component may be smaller or larger after adding communities.

6 Conclusions and discussion

In this paper, we have introduced the hierarchical configuration model, where the macroscopic graph is a configuration model, and on the microscopic level vertices are replaced by communities. We analytically studied several properties of this random graph model, which led to several interesting insights. For example, the condition for a giant component to emerge in the hierarchical configuration model is completely determined by properties of the macroscopic configuration model. However, the size of the giant component is also dependent on the community sizes. In contrast, the asymptotic clustering coefficient is entirely defined by the clustering inside the communities. For bond percolation on the hierarchical configuration model, the critical percolation value depends on both the inter-community degree distribution, and the shape of the communities.

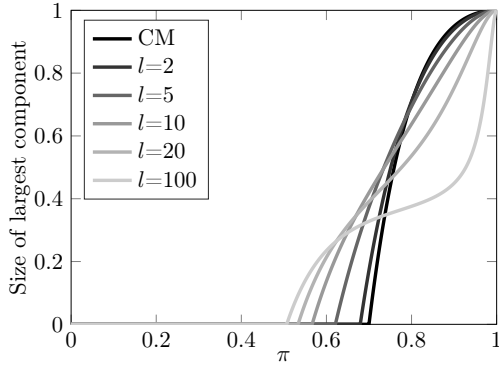


Figure 10: Size of giant component against π for line communities with different values of l

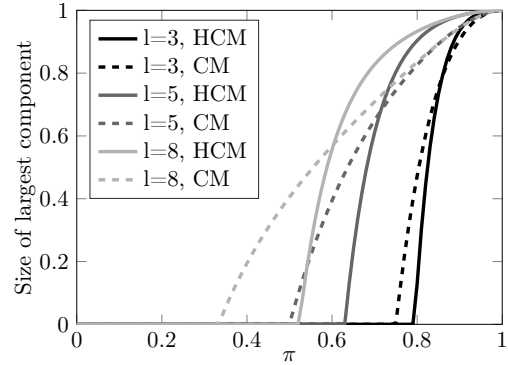


Figure 11: Size of giant component against π for star communities with different values of l

Furthermore, we have shown that if communities are dense with a power-law degree distribution, then the edges between communities follow a power-law with exponent that is one higher than the exponent of the degree distribution. We further investigate power-law relations in real-world networks, and compare these to the power-law relations in our hierarchical configuration model in a companion paper [17].

Finally, we have shown that several existing models incorporating a community structure can be interpreted as a special case of the hierarchical configuration model, which underlines its generality.

Worthwhile extensions of the hierarchical configuration model for future research include directed or weighted counterparts and a version that allows for overlapping communities.

The analysis of percolation on the hierarchical configuration model has shown that the size of the largest percolating cluster and the critical percolation value do not necessarily increase or decrease when adding clustering. It would be interesting to investigate how other characteristics of the graph like degree-degree correlations influence the critical percolation value.

Another interesting feature of the hierarchical configuration model is its applicability to real-world data sets. In this setting, the hierarchical configuration model creates a graph with the same degree distribution and the same community structure as real-world data sets.

Acknowledgement. This work is supported by NWO TOP grant 613.001.451 and by the NWO Gravitation Networks grant 024.002.003. The work of RvdH is further supported by the NWO VICI grant 639.033.806. The work of JvL is further supported by an NWO TOP-GO grant and by an ERC Starting Grant.

References

- [1] F. Ball, D. Sirl, and P. Trapman. Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. in Appl. Probab.*, 41(3):765–796, 09 2009.
- [2] F. Ball, D. Sirl, and P. Trapman. Analysis of a stochastic SIR epidemic on a random network incorporating household structure. *Mathematical Biosciences*, 224(2):53 – 73, 2010.
- [3] B. Bollobás. *Random Graphs*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2nd edition, 2001.
- [4] E. Coupechoux and M. Lelarge. How clustering affects epidemics in random networks. *Adv. in Appl. Probab.*, 46(4):985–1008, 12 2014.

- [5] M. Deijfen and W. Kets. Random intersection graphs with tunable degree distribution and clustering. *Probability in the Engineering and Informational Sciences*, 23:661–674, 10 2009.
- [6] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [7] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68:065103, Dec 2003.
- [8] R. van der Hofstad. Random Graphs and Complex Networks Vol. I. Available on <http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf>, 2015.
- [9] R. van der Hofstad. Random Graphs and Complex Networks Vol. II. Available on <http://www.win.tue.nl/~rhofstad/NotesRGCNII.pdf>, 2014.
- [10] R. van der Hofstad and N. Litvak. Degree-degree dependencies in random graphs with heavy-tailed degrees. *Internet Mathematics*, 10(3-4):287–334, September 2014.
- [11] S. Janson. On percolation in random graphs with given vertex degrees. *Electron. Journal of Probability*, 14:86–118, 2009.
- [12] S. Janson and M. J. Luczak. A new approach to the giant component problem. *Random Structures & Algorithms*, 34(2):197–216, 2009.
- [13] B. Karrer and M. E. J. Newman. Random graphs containing arbitrary distributions of sub-graphs. *Phys. Rev. E*, 82:066118, Dec 2010.
- [14] M. E. J. Newman. Random graphs with clustering. *Phys. Rev. Lett.*, 103(5):058701, July 2009.
- [15] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, Sep 2003.
- [16] P. Sah, L. O. Singh, A. Clauset, and S. Bansal. Exploring community structure in biological networks with random graphs. *BMC Bioinformatics*, 15(1):220, 2014.
- [17] C. Stegehuis, R. van der Hofstad, and J. S. H. van Leeuwen. Power-law relations in random networks with communities. 2015.
- [18] P. Trapman. On analytical approaches to epidemics on networks. *Theoretical Population Biology*, 71(2):160 – 173, 2007.